Scientometrics (2010) 82:647–662
DOI 10.1007/s11192-010-0177-9

Delineation of the genomics field by hybrid citation-lexical methods: interaction with experts and validation process

Patricia Laurens · Michel Zitt · Elise Bassecoulard

Received: 11 May 2009 / Published online: 13 February 2010

Abstract In advanced methods of delineation and mapping of scientific fields, hybrid methods open a promising path to the capitalisation of advantages of approaches based on words and citations. One way to validate the hybrid approaches is to work in cooperation with experts of the fields under scrutiny. We report here an experiment in the field of genomics, where a corpus of documents has been built by a hybrid citation-lexical method, and then clustered into research themes. Experts of the field were associated in the various stages of the process: lexical queries for building the initial set of documents, the seed; citation-based extension aiming at reducing silence; final clustering to identify noise and allow discussion on border areas. The analysis of experts' advices show a high level of validation of the process, which combines a high-precision and low-recall seed, obtained by journal and lexical queries, and a citation-based extension enhancing the recall. This findings on the genomics field suggest that hybrid methods can efficiently retrieve a corpus of relevant literature, even in complex and emerging fields.

Keywords Information retrieval · Bibliographic coupling · Genomics · Citation methods · Bibliometrics · Science mapping · Field delineation

Introduction

Analysis of leading edge, emerging or complex fields for science policy purposes is conditioned by the quality of the field delineation. Macro-level delineation, limited to the coarse-grain journal level, is a time and cost-saving methodology (see for example Rinia

P. Laurens (✉) · M. Zitt
Observatoire des sciences et des techniques (OST), 93 rue de Vaugirard, 75006 Paris, France
e-mail: patricia.laurens@obs-ost.fr

M. Zitt
e-mail: zitt@nantes.inra.fr

M. Zitt · E. Bassecoulard
INRA SAE2 LERECO—Unit 1134, Nantes, France
e-mail: bassecou@nantes.inra.fr

et al. 1993) that can be used to benchmark the production of actors within a given field. However, this crude Bradfordian selection does not achieve a good trade-off between silence and precision in complex and emerging fields. One needs to rely on the article level, instead of the journal level, in order to build a corpus of relevant publications. Adequate information retrieval techniques showing satisfactory levels of both recall and precision should be implemented.

Lexical queries are the traditional mode of delineation at fine-grain level. In the case of genomics, setting the search strategy is both cumbersome and limited in efficiency, because of complex field definition, fuzzy borders and high-tech content expressed by numerous technical acronyms likely to create noise. Expert's supervision is helpful to set up the queries, with the usual risks of specialization effects.

Complementary strategies built on bibliometric properties can be set up to overcome these limitations. Bibliometrics can be considered as the study of networks associated with scientific or technological activity: networks of scientists/institutions, of articles, of terms, of journals. Seldom does a bibliometric question receive an answer from a unique network: this is especially true for questions dealing with information retrieval, delimitation of fields, structuring and mapping of science. For example, the proximity between articles may be assessed from the authorship structure, relations of citations and/or share of common lexical contents. Garfield citation indexing (Garfield 1967) was in this respect a revolution, breaking the monopoly of lexical forms of retrieval.

The time has probably come to promote combined and/or hybrid techniques, together with competing solutions for addressing a particular issue, to try and gain advantage from this multifaceted reality. Google, the most famous retrieval system worldwide, is indeed a hybrid system (Brin and Page 1998), combining lexical search with hyperlinks - quasi-citation linkages. In citation bibliometrics, hybridization also opens up a wide space to design efficient retrieval processes.

Here an application of a combined lexical and citation process for field delineation is described (in the followings referred to as the "lex + cite" method), basically drawn from Zitt and Bassecoulard (2006). The method was applied to the retrieval of publications in the field of genomics and was closely supervised by experts throughout the successive steps. The present article focuses only on the application of the delineation issue to this particular field.

The first step of the delineation process is a bibliometric mix of classical methods, a (drastic) selection of core journals and the construction of a lexical query. It ended at the harvest of a first set of relevant publications. The dataset was then enhanced by a citation-based extension technique. Lastly, a final clustering on lexical content was performed to refine delineation and discuss border areas.

A panel of experts, scientists from various subfields of genomics (see acknowledgements), has been set up to monitor the entire process and to provide assistance at different key point steps. The panel was involved in the selection of the core genomics journals and the formulation of the lexical query using genomics related keywords, the assessment of the citation-based extension process and the final proposals of noise reduction on a map and analysis of lexical clusters.

After recalling the overall context of the present study (second section), the paper will shortly present the different steps of the "lex + cite" method with the contribution of experts to the process and/or to the validation. Because of this structure, the sections follow the different steps of the process, presenting for each the method used, together with the meta aspects of experts involvement: third section is devoted to the lexical initialization,

fourth section to the citation-based extension, fifth section to the noise-reduction using a clustering stage, before discussion and conclusion.

Context: the CSTG (sciento-technology corpus in genomics) project and the definition of the genomic field

The process of field delineation described here is carried out within a larger project funded by the french National Research Agency (ANR) on the building of databases for social scientists, especially economists, in the field of genetics, with a particular focus on genomics.[1] It is dedicated to the analysis of two separate methods of diffusing information about human genetic sequences: through patents in technology on the one hand, through scientific publications in science on the other hand. The aim of the project is to investigate the relationship between science and technology in this particular field and to study the characteristics of this newly established R&D field. The present paper only addresses the bibliometric work package of the project, the delineation issue on the publications side.

For publications, the database was defined as extracts from Thomson-Reuters's Web of Science (WoS). The EPO Worldwide Patent Statistical Database (PATSTAT) was used on the patents side, for memory sake. As various exploitations are planned, two quite distinct delineations were carried out. The first one, very coarse, concerns the vast area of genetics, for the purpose of analyzing co-activity publications-patents especially. Both WoS and PATSTAT have been searched using a similar search strategy involving genetics and genomics keywords. The keyword list is inspired from the Georgetown University query,[2] with some add-ons of specific genomics-related terms.

This paper details the second delineation process, focused on the sub-area of *genomics*, where fine-grain network analysis was conducted. How does one define "genomics"? Both genetics and genomics involve gene-related research. The term "genomics" was coined in 1986 when launching the eponym new journal (McKusick and Ruddle 1987). Genetics *stricto sensu* is the science of gene heredity and variation of organisms by looking at single genes, one at a time, as a snapshot. In contrast, genomics typically looks at all the genes or at least at large fractions of a genome as a dynamic system, over time, to determine how they interact and influence biological pathways, networks and physiology, in a much more global sense. It encompasses everything from sequencing genomes, ascribing functions to genes, and studying the structure of genes (gene architecture). The most famous project in this field was the worldwide project dedicated to the sequencing of the Human genome in the period 1990–2003 (IHGSC 2004).

Especially since 2003, a new subfield of genomics has emerged, namely post-genomics, which focuses on the understanding of biological phenomena involving genes, gene functions and genes-related products. Post-genomics uses techniques previously developed for genomics and takes them further, studying patterns in gene transcription to form proteins (transcriptomics), in genes expression as proteins (proteomics), and in genes influence on the chemicals that control the cellular biochemistry and metabolism (metabolomics).

[1] Consortium: ADIS (Université Paris-Sud), Lereco (INRA), OST.

[2] See http://dnapatents.georgetown.edu/.

Building the initial set

To start the process, experts were asked to give their general view of the field of genomics, its relation to genetics and border areas. They stressed the consensual following points, echoing the above mentioned milestones and allowing the specification of genomics literature:

- genomics involves a "large scale" approach of genes or genes related products and the study of those genes in their complex biological environment.
- post-genomics publications should be included in a genomic publication corpus
- bioinformatics is included as well as research devoted to the development of specific instrumentation for genomic large scale purposes.

Besides these selection rules, the panel stressed that since genetics and genomics fields are strongly interlinked, some overlaps occur and the attribution of a particular article to either genetics or genomics may be difficult.

For the initial input, bibliometric teams with an experience of micro-level studies report a variety of combinations (for example, Debruin and Moed 1993; Aksnes et al. 2000; Van Leeuwen et al. 2001; Bassecoulard et al. 2007), often based on experts' advice: selection of journals, key-words, key-authors or institutions, or key-cited works. Here we choose a combination of core journals and lexical query. The source of primary data is the OST off-line version of the Thomson-Reuters SCI-Expanded Web of Science (WoS), for database years 1999–2005. Only articles and letters have been kept.

Specialized journals

Our starting point was a list of core journals in genomics. The identification was found to be rather tricky, because of the imbrication of genomics and genetics. Many journals titles contained the term "gene" but, in the view of the panel, did not belong to "genomics". Among a set of 60 pre-selected journals, experts retained a restricted list of 26 core journals (see Annex I), belonging to the following Thomson-Reuters subject areas: 'Genetics and Heredity', 'Biochemistry & Molecular biology', 'Biotechnology & Applied Microbiology' or 'Cell biology'. Some divergences were expressed on journals such as 'DNA Repair' and 'DNA Research'.

The list of journals selected by the panel can be compared to that selected by Basu and Lewison (2006) from the free access Pubmed database for the same field. From the 24 journals selected by these authors, 14 are in the WoS database and were selected by our panel. All the articles and letters from our core journals set are retained in the corpus. This corresponds to ca. 15,900 documents.

Lexical query

The formulation of a lexical query filtering genomics out of genetics is a challenging task since the same biological materials are involved in both fields. Few previous studies reported genomics field delineation using a lexical query approach (Archambault et al. 1999).

The experts were asked to select keywords specific to genomics literature. They rejected the keywords related to biological materials: *sequence*, *antisense*, *yac*, *bac*, *nucleoside*, *haplotype*, *exon*, *intron*, *nucleotide*, *nucleic acid*… These terms were deemed non specific of genomics, except *genome* and *snp* (for *single nucleotide polymorphism*). Terms such as *transcriptomics, pharmacogenomics, proteomics, bioinformatics* and *genebank* were

Table 1 Lexical formula

((__dna or _dna or dna or __rna or _rna or rna or rna_ or dna_ or _rna_ or _dna_ or __dna_ or __rna_ or gene or genes or genet% or genot% or chromoso%)	*Standard genetic terms*
AND	
(large_scale or (large scale) or high_throughput or high_throughputs or (high throughput) or (high throughputs) or (expression profile) or (expression profiles) or librar% or sequencing or screening or shuffling or synten% or map%))	*Large scale terms*
OR	
(%genom% or genebank or genebank_ or metabolom% or transcriptom% or pharmacogenom% or toxicogenom% or epigenom% or proteom% or interactom% or metagenom% or omic or omic_ or bioinformatic% or snp or snps)	*Specific genomic terms*

selected, but are too general to be used alone. A Boolean lexical query using only these generic terms retrieved less than 6,000 publications between 1999 and 2005. Adding combinations of standard genetic terms like *DNA*, *RNA*, *gene* or *chromosome* with terms qualifying the large-scale approach of genomics, proved successful. The final query is shown Table 1.

Experts were also asked to validate target lexical fields. The outcomes of a lexical query vary in magnitude depending on the lexical fields where it is applied. Classically, databases offer all or some fields amongst the following: titles; authors' keywords; database keywords, sometimes of several kinds; abstracts; full text. The WoS offers all but the last one, and the database keywords are quite original (keywords+), resulting from a hybrid analysis (through cited items). As they are likely to be noisy in our context, they were not used.

Three strategies were compared: (a) on titles; (b) on titles + authors' keywords; (c) on titles + abstracts. The three ways were tested with the above Boolean formula, and experts discussed samples of queries outcomes. The outcomes of the strategies (a) and (b) were found relevant by the experts. Strategy (c) resulted in quite a large set, which was severely rated by the panel because of a large dispersion out of genomics. The strategy (b) was eventually selected since it increased the retrieval rate by 16% without evidence of increasing the noise. Keeping a low level of noise was considered a key point at this stage since the next step, the citation stage, aimed mainly at reducing silence, a point discussed further.

The lexical query added ca 36,600 documents to those coming from the core journals. The resulting initial set (union core journals + lexical query) contains ca 52,500 documents. This set was used as the seed literature to further collect an extension based on citations links between articles.

Citation based extension

Informetric process

The principle of the citationist extension has been described by Zitt and Bassecoulard (op.cit. 2006) and further commented on in Bassecoulard et al. (op.cit. 2007). The aim is to extend the initial set of literature, (the seed literature), to the "sister literature", i.e. publications sharing the domain-specific knowledge base of the initial set. For this purpose it is assumed, in a Mertonian interpretation, that the set C of articles cited by the seed is a proxy of the knowledge base.

The protocol consists in recalling new relevant publications (outside the seed) that cite this reference literature. The global rationale is bibliographic coupling, but instead of computing coupling links at the document level—a rather heavy task in terms of computing requirements—we considered the current initial set of literature as a macro-document. A particular advantage thereof is that articles may be retrieved without having much proximity with any particular article in the current set as soon as they contain any combination of structuring items (here the specific cited articles). A shortcoming is that the corpus size affects the process, hence the desirable setting of the parameters.

The general procedure involves three parameters, further combined into synthetic relevance measures.

– two parameters on the cited side (for each article cited by the seed) to qualify the domain-specific reference literature

y^i, *genericness*, is the "local" citation score, i.e. the citations retrieved from the seed (m citing articles, $j = 1...m$) for the i^{th} cited article in the cited set C:

$$y^i = \Sigma_{j(1...m)} c^i{}_j \text{ where } c^i{}_j = 1 \text{ if i is cited by j, 0 otherwise}$$

Let Y a threshold on y^i. The higher Y, the most selective we want to be, discarding marginal cited items and thematic areas.

u^i, *specificity*, is the ratio of y^i, local citations—received from the seed—and y'^i global citations—received from the whole Web of Science (WoS, M citing articles $h = 1...M$):

$$u^i = y^i \Big/ y'^I \text{with } y'^i = \Sigma_{h(1...M)} c^i{}_h$$

where $c^i{}_h = 1$ if i is cited by h, 0 otherwise. A corresponding probabilistic form for specificity is obtained by dividing by the expected value:

$$v^i = u^i/(m/M).$$

Let U a threshold on u. The higher U, the most specific we want to be, by discarding cited articles with little specialization towards the field.

After the application of the two thresholds Y and U on cited articles from C, we obtain a subset of domain-specific cited articles. For convenience, let us consider D ($g = 1...p$), the domain-specific subset of C.

– one parameter on the citing side to retrieve the domain-relevant "sister literature"

Let NSR(k) (number of specific references) the number x^k of references that a citing article k from the WoS ($k = 1...M$) counts in the set D of domain-specific cited articles:

Let z^k the total number of references of article k.

$$\text{NSR}(k) = x^k = \sum\nolimits_{f(1...z^k)}^{k} r^k{}_f$$

where $r^k{}_f = 1$ if the reference f is domain-specific and falls in D, 0 otherwise

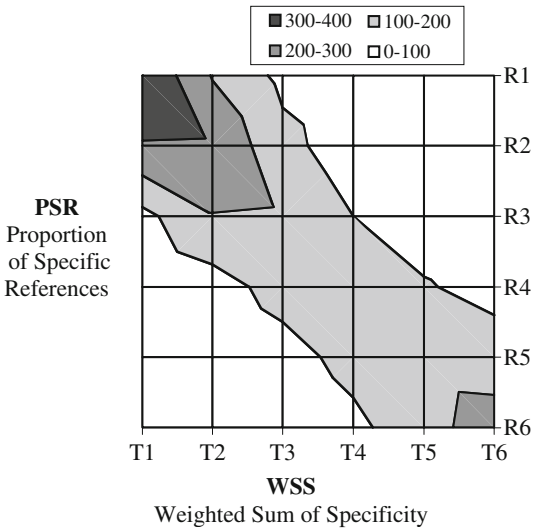

Fig. 1 Probabilistic indexes on Table 2

NSR(k) is a proxy of total relevance. Let X the threshold on NSR(k).

An alternative measure of total relevance, used here, is:

$WSS(k) = \sum^{k}_{f(1...z)} u^{i}_{f}$, the weighted sum of the specificity index over all references of the citing article k. u^{i}_{f} denotes the specificity of the cited article i corresponding to the reference f in article k. $u^{i}_{f} = 0$ if the article i does not fall in C.

Note that WSS is less sensitive to a threshold on U than x^k, and can be calculated without U threshold.

In addition to total relevance, likely to depend on the total number of references, let us now address the relative relevance. A proxy of relative relevance is:

$PSR(k) = x^k/z^k \leq 1$, the proportion of specific references among the total references of article.

– An alternative measure is the weighted proportion of references:

$$WPR(k) = WSS(k)/\ z^k \leq 1.$$

With its multistage arrangement, the extension process is a bit different from a standard retrieval scheme, but the global rationale has some analogy with the tf-idf approach (Fig. 1).

In the protocol used here, thresholds Y, U and X were used beforehand at a low selective level, in order to alleviate the computer requirements. In the present application, the presets were Y = 2, U = 0.2, giving ca. 181,500 domain-specific articles that receive at least 20% of their citations from the initial set. They can be considered as the elements of a Boolean citation-based query, retrieving 456,400 documents, among which 406,400 do not belong to the initial set. On the citing side the preset is X = 3. Citing documents with at least 3 domain-specific references were kept, an additional filter on document types was

Table 2 Cross-tabulation of the filtered citation-based extension retrieved articles ranked according to two proxies of relevance

WSS quantiles*	Up to 10%	10 to 20%	20 to 40%	40 to 60%	60 to 80%	80 to 100%	Total
PSR quantiles	T1	T2	T3	T4	T5	T6	
Up to 10%	5414**	1892	1466	602	300	121	9795
R1	0.67–8.24	0.65–3.69	0.64–2.34	0.65–1.47	0.62–1.05	0.64–0.78	0.66–5.75
10 to 20%	2850	2819	2526	899	418	257	9769
R2	0.40–6.31	0.39–3.63	0.39–2.34	0.38–1.47	0.38–1.06	0.37–0.79	0.39–3.69
20 to 40%	1385	3810	7273	3910	2047	1137	19562
R3	0.26–5.55	0.25–3.53	0.24–2.28	0.23–1.47	0.23–1.06	0.23–0.77	0.24–2.38
40 to 60%	114	1076	5440	5863	4234	2622	19349
R4	0.16–5.30	0.16–3.41	0.15–2.17	0.15–1.45	0.15–1.05	0.15–0.77	0.15–1.61
60 to 80%	12	165	2343	5404	6260	5742	19926
R5	0.10–5.94	0.11–3.31	0.10–2.08	0.10–1.43	0.10–1.04	0.10–0.77	0.10–1.21
80 to 100%	7	20	516	2885	6584	9403	19415
R6	0.07–5.44	0.06–3.49	0.07–2.01	0.07–1.40	0.06–1.03	0.06–0.76	0.06–0.98
Total	9782	9782	19564	19563	19843	19282	97816
	0.53–7.25	0.35–3.57	0.24–2.23	0.17–1.44	0.12–1.04	0.10–0.76	0.22–2.18

WSS weighted sum of specifivity, *PSR* proportion of specific references

* Non cumulative classes of percentiles from the top value of the variable through the bottom value

** In each cell, first row: number of documents. Second row: mean value of PSR–WSS within the cell

practiced. This operation finally left 97,800 articles and letters in the filtered citation-based extension, submitted to expert assessment. These successive sets are shown on Fig. 3.

The set of retrieved articles was ranked by relevance based on two bibliometric parameters. The first one is a proxy for relative relevance, the proportion of specific references (PSR); the second one a proxy of total relevance, the weighted sum of the specificity (WSS). A cross-table was established (Table 2) and the articles were dispatched into one of the 36 cells. For convenience, we used "quantiles" (not cumulative) categories; 10, 10, 20, 20, 20, 20% denoted T1–T6 on WSS and R1–R6 on PSR.

The first thing we learn from this table is the relation between the absolute and relative measure, illustrated by the probabilistic index on the table[3], with all high values alongside the diagonal.

Testing the relevance of the bibliometric ranking was then performed, by confronting it to the judgment of experts. Two outcomes were expected from the experts:

(a) validation of the bibliometric (relevance) ranking of articles added by the extension, namely the ranking of groups of articles represented by a cell in the Table 2. A satisfactory match between the bibliometric ranking and the experts' ranking would be a strong argument in favour of the soundness of the process. For this purpose, experts were asked to qualify, on samples drawn from each cell (sample size: 40 articles) of the cross-table, each document as belonging or not to "genomics", thus allowing to assess on each cell, on a statistical basis, the proportion of relevant items.

[3] Probabilistic index = observed cell population c(ij)/expected cell population where expected cell population = c(i.)c(.j)/c(..).

Table 3 Experts judgement on a sample of the extension: % of genomic documents per cell

WSS PSR	Up to 10% (T1) (%)	10 to 20% (T2) (%)	20 to 40% (T3) (%)	40 to 60% (T4) (%)	60 to 80% (T5) (%)	80 to 100% (T6) (%)
Up to 10% (R1)	90.6	83.1	86.6	73.6	57.4	76.6
10 to 20% (R2)	76.5	73.4	74.0	68.0	63.9	63.0
20 to 40% (R3)	71.1	65.4	60.4	54.8	54.4	48.0
40 to 60% (R4)	68.0	61.1	51.5	56.5	40.0	35.6
60 to 80% (R5)	ns	50.5	45.6	43.8	39.7	31.4
80 to 100% (R6)	ns	ns	32.8	35.0	18.6	27.8

ns: number of the documents in the cell lower than 20

The question is then: do the results confirm, first at the expert group level, then at the individual expert level, the ranking by bibliometric relevance.

(b) tuning of the citation-based extension. As in information retrieval processes based on Bradfordian rules, no "natural border" can be defined for most scientific fields. Bibliometric relevance assessment and information retrieval (IR in the followings) trade-off of experts helped to reach a practical solution. We limited ourselves to technical aspects, without external cost functions. Experts were asked to rate each cell as globally acceptable or not, in other words to reveal their own IR trade-off.

Tuning the extension process

Table 3 shows the results of the expertise at an aggregate level by giving the percentage of articles relevant to genomics for each cell. For either measure (PSR or WSS), the average percentage of genomics-related articles decreases when the measure decreases. This is also true for conditional distributions to one class, either row-wise or column-wise, with very few exceptions. Interestingly, the dispersion of experts' individual visions of the field, restricted or generous, proved quite large. By the way, this source of variation is much stronger than the sampling process itself. For example, ratings in cell R1-T1 varies from 77 to 100%, in cell T2-R2 from 52 to 95% etc. Nevertheless, resulting cell rankings were alike and globally matched the trends of bibliometric measures (decrease from column T1 to column T6, and from row R1 to row R6). This is a strong indication in favour of the soundness of the citation-based extension.

If we look into details, we can visualize the comparison within PSR and WSS from figure (Fig. 2). The relative measure performs better on high deciles cells with non significant values (bottom left) are conventionally represented by negative values (dashes).

The revealed individual IR trade-off of experts was also diverse. We eventually used a conservative strategy and only publications from cells considered as "mainly non genomics" by all experts were discarded from the extended set (dashed cells). These options yielded a final citation-based extension of ca. 67,200 documents. The outcome of the various stages is sketched in Fig. 3.

The final set has ca. 119,700 documents (union of initial set and final citation-based extension), 56% of them being retrieved by the citation extension step. We will come back to the IR features of the citation-based process in the next section.

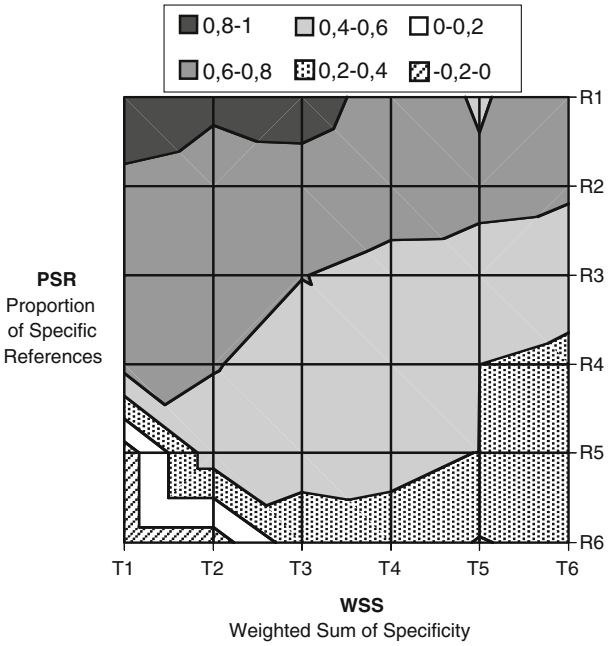

Fig. 2 Experts judgement on a sample of the extension: % of genomic documents per cell

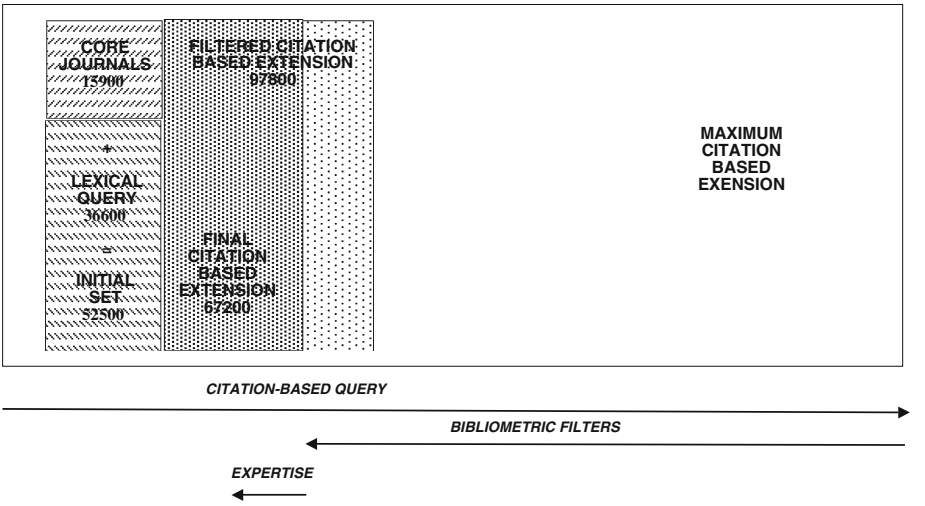

Fig. 3 Successive sets of the hybrid process

Mapping and clustering

The final stage of clustering on lexical content is meant to feed discussion of border areas, or of areas appearing as "clustered noise", out of the topic. Many clustering algorithms are available. Here we used the axial K-means method (AKM, Lelu and François 1992;

Table 4 Experts judgement on clusters according to their relevance to genomics and the relative weight of the extension in the cluster

Share of the citation-based extension	Core genomics	Border theme	Non genomics	Total
Low (≤44%)	7	–	–	7
Average	21	3	10	34
High (≥71%)	6	2	1	9
	34	5	11	50

Population of clusters calculated on primary assignments of documents (no overlaps)

Lelu 1994) in a rationale of lexical coupling. AKM is an axial variant of the well-known K-means clustering algorithm[4] and is able to process large volumes. The first step is a K-means partition, each document is assigned to one cluster. In a second step, projections of documents upon the axoid of each cluster allow multi-assignment of documents. Principles are recalled in Bassecoulard et al. (op. cit. 2007), in an application to nanosciences. A hybrid clustering on citations and lexical content has been reported elsewhere, again on a nanosciences corpus (Zitt et al. 2008), a similar exercise is on-going on the present corpus.

Term-extraction was carried out on titles and abstracts of the final dataset. The inclusion of abstracts at this stage is not as risky as it was at the delineation stage. We set the breakdown at 50 thematic clusters. Experts proposed denominations for clusters, after an analysis of the titles of top assigned articles and terms. They also qualified the global relevance of each cluster for genomics into three groups: Core genomics; Border theme; Non genomics.

Experts validated 34 clusters out of 50 as Core genomics clusters (human genome, plant genome polymorphism, bioinformatics, proteomics etc.). 5 clusters were assessed as border themes (two clusters devoted to viruses) and considered as promising, though not yet central. For the remaining clusters, experts emphasized the risk of noise, as topics were more focused on the single-gene level or on methods not specific to genomics. Themes and corresponding qualifications are listed in Annex II.

Extension process and field borders

This question was investigated by several means. First the proportion of papers from the citation-based extension in each of the 50 clusters was calculated and compared to its share in the whole dataset (56%). According to their citation extension share, the 50 clusters were classified as 'Low' (less than 44% of the cluster's publications were retrieved during the citation extension step), 'Average' (45 to 70% publications retrieved from the extension step) and 'High' (more than 70% of the publications retrieved from the extension phase). We crossed these classes with the classes of global relevance assessed by the experts (Table 4).

As expected, the seven clusters with a high contribution in the initial set belong to core genomics. For the nine clusters with a high contribution of the extension, one of them is considered as noisy, two others as border themes, but six of them, the majority, were also assigned to core genomics. The extension process brings some new topics along and also enhances central ones.

[4] AKM is implemented in the commercial software Neuronav by Diatopie (S. Aubin, www.diatopie.com). It is enhanced it with a basic but robust and efficient term-extracting sequence.

For each "retrieval set" (core journals, addition by the lexical query, citation-based extension), we analyzed the distribution of documents between the three groups of clusters, classified according to their global relevance towards genomics. As shown in Table 5, the percentage of documents that are primarily assigned to core genomics clusters decreases moderately, from core-journal documents (80.7%) to articles added by the lexical query and articles from the citation-based extension (68.6%). There is a shift towards documents assigned to border themes. For documents assigned to noisy clusters, the proportion is almost the same for articles added by the lexical query (outside core journals) and for documents brought by the citation-based extension process.

Finally, we checked the primary assignments of documents of the extension displayed by bibliometric criteria, as in the ex-ante assessment on cell samples. Results shown in Table 6 corroborate previous judgments on samples from Table 3, with a 0.86 correlation on cell rankings.

When comparing the addition brought by the lexical query (to the core journals), and the further extension brought by the citation process, we could expect the latter to be much noisier. In fact, the quality of the citation-based extension, in terms of types of clusters (core, border, out), appears remarkable. Now, if we go back to the tuning of the extension, following experts' advices ("Citation based extension" section) and especially their choice of cells to discard, we observe that the experts did not demand a very high level of precision, perhaps as a result of the difficulty to distinguish between genomics and genetics publications in practice. We have also seen that the perception of genomics could be quite different amongst them. The precision on the final extension can be approximated from the number of documents and the proportions estimated on each cell (Tables 2 and 3), at ca. 60%. Since it was too heavy to have a validation of documents outside the "filtered

Table 5 Distribution of documents from retrieval sets by relevance groups of clusters

Global relevance of clusters (experts judgment)	Core journals (%)	Addition from Lexical query (%)	Total initial dataset (%)	Final citation-based extension (%)	Total final dataset (%)
Core genomics	80.7	73.2	75.4	68.6	71.6
Border theme	4.1	6.7	5.9	9.5	7.9
Non genomics	15.2	20.1	18.7	21.9	20.5
Number of documents	100	100	100	100	100
	15900	36600	52500	67200	119700

Percentages calculated on primary assignments of documents (no overlaps between clusters)

Table 6 Experts judgment on the extension: % of documents assigned to core genomics clusters

WSS PSR	Up to 10% (T1) (%)	10% to 20% (T2) (%)	20% to 40% (T3) (%)	40%to 60% (T4) (%)	60% to 80% (T5) (%)	80% to 100% (T6) (%)
Up to 10% (R1)	85.2	81.9	80.6	79.7	78.0	74.4
10 to 20% (R2)	77.3	74.6	74.0	71.9	66.0	70.8
20 to 40% (R3)	72.9	70.4	68.4	65.7	64.7	61.6
40 to 60% (R4)	71.1	65.4	63.4	62.2	62.2	
60 to 80% (R5)	ns	73.9	63.5	58.8		
80 to 100% (R6)	ns	ns	64.0	60.0		

Percentages calculated on primary assignments of documents (no overlaps between clusters)

citation-based extension" analysed in the tables above, the recall could only be very roughly approximated, by extrapolating the trend observed on WSS (or PSR), using a linear model, the only compatible in the available range of data. On theses bases, the revealed recall would be about 3/4, likely to be overestimated by the linear extrapolation. These figures should be taken as large approximations. In any case, the bibliometric tables could be a guideline for a variety of other trade-offs.

In order to establish the "lex + cite" method as a robust and realistic production tool for field delineation, without engaging heavy expert involvement in run-of-the-mill applications, the relevance analysis could be limited to a "critical zone" (here rows R5-R6 * columns T5-T6 in the tables) where bibliometric parameters become lower as those encountered in the lower fraction of the initial seed, the fraction being defined on adequate groups of publications. This is possible if a high precision is warranted for the seed, as mentioned above. Also, the test on cross-tabulated cells based on two indices (here PSR and WSS) can be avoided by picking a single measure, either among NSR, WSS, PSR, WPR or various retrieval formulas in line with tf-idf. The last one, WPR, could be promising. Moreover, pre-selections at sensible thresholds can be helpful to alleviate the computing process. Using a single index and restricting the expert validation to a critical area seems a sensible way for practical implementations of the "lex + cite" method.

Conclusion

Three main results were obtained in this test of "lex + cite" methods. Firstly, in the case of areas where generic terms cannot be avoided, such as genomics, lexical queries—at least in Boolean form—are explosive when applied to long target fields like abstracts. In these particular areas, it is strongly recommended to combine a lexical query on short fields (e.g. titles words and authors' keywords) and a citation-based extension to achieve a good IR trade-off. The range of application of hybrid delineation is not limited of course to those particular situations. Secondly, a systematic validation of the citation extension process was conducted with the help of experts. Experts showed a large individual variety in their perception of genomics as a subfield of genetics, but their assessment of the successive sets of the extension proved remarkably consistent with the bibliometric process grounded in relevance assessment. Thirdly, as expected, the clustering stage proved a good tool to feed experts' discussion about the borders of the topic, which validated the "lex + cite" process from another point of view. The citation extension process contributes to the enhancement of themes present in the initial lexical set. It also brings new themes, most of them considered as relevant by experts. In advanced methods of delineation/mapping, hybrid methods appear as a promising path to capitalize the advantages of approaches based on either words or citations, and possibly other bibliometric relations. In this experiment, experts were involved in a systematic process that proves efficient, both to validate the principle and to tune the delineation sequence. The lessons from the current experiment will help to alleviate the experts' validation process in future studies.

Aknowledgments This work is part of the CSTG project launched by Antoine Schoen and Bertrand Bellon, a project supported by ANR. The authors are indebted to Sylvain Aubin (DIATOPIE) for term extraction and clustering. They also thank the panel of experts for their commitment: P. Bessières (INRA), A. Lecharny (CNRS-Evry), M. Pinto (Université Paris XI), MM JP Rousset and M Dubow (IGMORS-Université Paris XI), P. Wincler (Genoscope). We also benefited from the experience in the field of Bérangère Virlon (OST). The authors are solely responsible for the views expressed in this article.

Annex I List of core journals

ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS	GENOME
BIOINFORMATICS	GENOME BIOLOGY
BMC BIOINFORMATICS	GENOME RESEARCH
BMC GENOMICS	GENOMICS
BRIEFINGS IN BIOINFORMATICS	JOURNAL OF PROTEOME RESEARCH
COMPARATIVE AND FUNCTIONAL GENOMICS	MAMMALIAN GENOME
CURRENT GENOMICS	MOLECULAR & CELLULAR PROTEOMICS
CYTOGENETIC AND GENOME RESEARCH	MOLECULAR GENETICS AND GENOMICS
DNA REPAIR	PHARMACOGENETICS AND GENOMICS
DNA RESEARCH	PHARMACOGENOMICS
DNA SEQUENCE	PHARMACOGENOMICS JOURNAL
EXPERT REVIEW OF PROTEOMICS	PHYSIOLOGICAL GENOMICS
GENES CHROMOSOMES & CANCER	PROTEOMICS

Annex II List of themes

The clusters were labelled arbitrarly from M1 to M50 during the clustering process. The most central terms related to each cluster are shown to point out its thematic content

Core genomics

Extension share: high	
M18/Population_genomics	M32/Marker/RAPD/AFLP/Polymorphism
M20/Resistance/Resistance_genes/ Plant_&_Trout_resistance	M40/QTL/Trait/Mapping/Polymorphism
M31/LOH/Tumor_suppressor/Genome_&_Cancer	M47/Species/Phylogeny/Evolutionary_genomics
Extension share: average	
M 3/Plant_genomics/Transgenic_plants	M25/Patient/Disease_genomics/Biomarkers/ Pharmacogenomics
M 4/DNA_sequence/Satellite	M27/Evolution/Evolutionary_genomics
M 5/Strain/Microbial_genomics	M28/Cancer/Genome_&_cancer
M 6/Cell_identity_&_Gene_expression	M35/C-DNA/Transcription/C-DNA_library
M 8/Alignment/Bioinformatics	M36/Polymorphism
M12/Network/Biological_networks/Model	M43/Mouse/Murine_genomics
M14/Locus/Microsatellite_locus/Polymorphism	M44/Expression/Cell_identity_&_Gene_expression
M15/Cell_line/Tumor/Genome_&_Cancer	M45/LOD/Linkage_analysis/Polymorphism
M16/Spectrometry/Proteomics	M46/Human/Primate/Gene_annotation/ Comparative_genomics
M22/Human/C-DNA/Gene_annotation	M48/C57BL/Congenic_strains/Murine_genomics
M23/Exon/Genomic_organization/Gene_annotation	
Extension share: low	

Appendix continued

M 1/Human_genome/Human_genome_project	M17/Map/Linkage_maps/Polymorphism
M 9/Genome/	M24/System/Systems_biology/Bioinformatics
M10/Comparative_genomic_hybridization/Tumor	M38/Genome/Genome_sizes
M11/SNPs/Polymorphism	

Border themes

Extension share: high	
M 2/Translocation/FISH/leukemia	M21/Hybrid/Somatic_hybrids/Ferility
Extension share: average	
M13/Transcriptional/Saccharomyces_cerevisiae/ Transcriptome	M50/Virus/Virus_replication/ Virus_recombinatiio,
M26/Virus/Nucleotide_sequence	
Noisy, mostly Non genomics	
Extension share: high	
M30/Mutant/Mutagenesis	
Extension weight: average	
M 7/Enzyme/Escherichia_Coli	M37/Cell/DNA_damage
M19/Repair/DNA_damage	M39/DNA/Arrays/Genomic techniques
M29/Promoter/Transcription	M41/Signaling/Kinase/MAPK
M33/RNA-/Virus	M42/Mutation/Missence_mutation
M34/PCR/Methods/applications	M49/Residue/Amino_acid_sequence

References

Aksnes, D. W., Olsen, T. B., & Seglen, P. O. (2000). Validation of bibliometric indicators in the field of microbiology: A Norwegian case study. *Scientometrics, 49*(1), 7–22.

Archambault, É., Gingras,Y., Godin, B. & Vallières F. (1999). Characterization of genomics in Canada—a bibliometric study of scientific articles and research grants 1995–1997. Prepared for Genome Canada by OST. 19 pp.

Bassecoulard, E., Lelu, A., & Zitt, M. (2007). Mapping nanosciences by citation flows: a preliminary analysis. *Scientometrics, 70*(3), 859–880.

Basu,A., & Lewison, G. (2006). *Visualization of a scientific community of Indian origin in the US: A case study of bioinformatics and genomics*. International Workshop on Webometrics, Informetrics and Scientometrics & seventh COLLNET meeting, 10–12 May 2006, Nancy.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and Isdn Systems, 30*(1–7), 107–117.

Debruin, R. E., & Moed, H. F. (1993). Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. *Scientometrics, 26*(1), 65–80.

Garfield, E. (1967). Primordial concepts, citation indexing and historio-bibliography. *Journal Library History, 2*, 235–249.

International Human Genome Sequencing Consortium (IHGSC). (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931–945.

Lelu, A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday & Y. Lechevallier (Eds.), *New approaches in classification and data analysis* (pp. 241–248). Berlin: Springer-Verlag.

Lelu, A., & François, C. (1992). Automatic generation of hypertext links in information retrieval systems. In D. L. al (Ed.), *Proceedings of ECHT'92 (Milano)* (pp. 112–121). New York: ACM Press.

McKusick, V. A., & Ruddle, F. H. (1987). A new discipline, a new name, a new journal. *Genomics, 1*(1), 1–2.

Rinia, E. J., Delange, C., & Moed, H. F. (1993). Measuring national output in physics—delimitation problems. *Scientometrics, 28*(1), 89–110.

van Leeuwen, T. N., van der Wurff, L. J., & van Raan, A. F. J. (2001). The use of combined bibliometric methods in research funding policy. *Research Evaluation, 10*, 195–201.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management, 42*(6), 1513–1531.

Zitt M., Lelu A., & Bassecoulard E. (2008). Hybrid maps of scientific fields (terms and citations): an application to nanosciences. In J. Gorraiz & Schiebel, E. (Eds.), *Excellence and emergence. A new challenge for the combination of quantitative and qualitative approaches. 10th International conference on science & technology indicators. Book of abstracts*. Vienna, Austria, 17–20 September, 2008 (pp. 53–56). Vienna (AUT): Austrian Research Centers GmbH.